Storage Research at ORNL

Presentation to HEC-IWG Workshop

Sudharshan Vazhkudai

Network and Cluster Computing, CSMD

R. Scott Studham

National Center for Computational Sciences

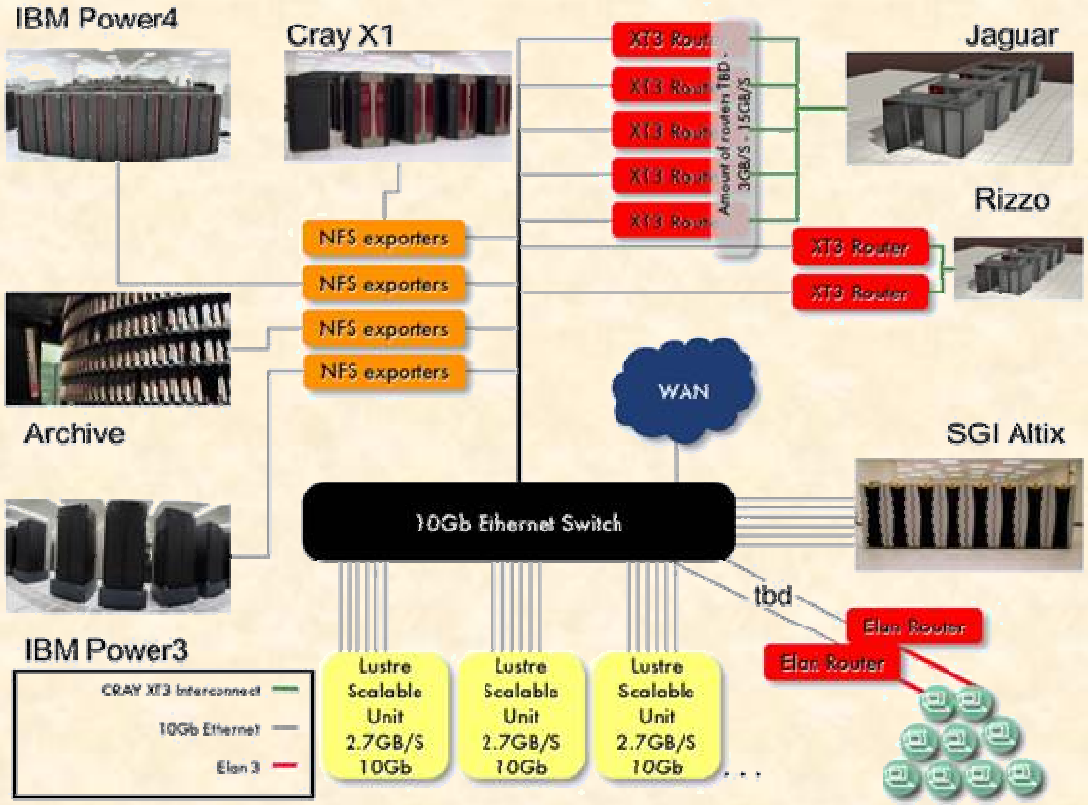
Contributors: Sarp Oral, Hong Ong, Jeff Vetter, John Cobb, Xiaosong Ma and Micah Beck

Application Needs and User Surveys—Initial Observations

- GYRO, POP, TSI, SNS
- Most users have limited IO capability because the libraries and runtime systems are inconsistent across platforms.
- Limited use of Parallel NetCDF or HDF5
  - POP moving to P-NetCDF
  - SNS uses HDF5
- Seldom use of MPI-IO
- Widely varying file size distribution
  - 1MB, 10MB, 100MB, 1GB, 10GB

Current Storage Efforts for NLCF

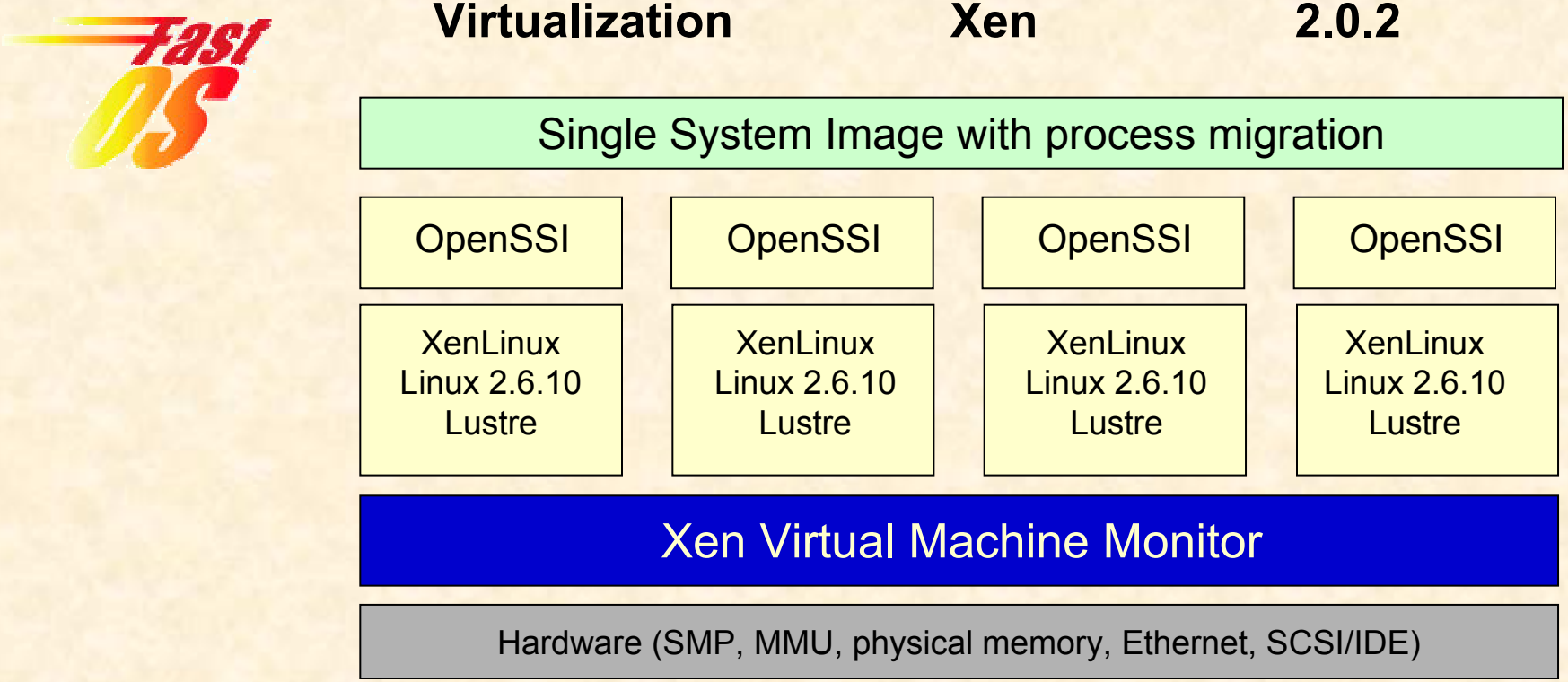
- Future procurements require support for center–wide file system
- Minimize the need for users to move files around for post processing.
- As most applications continue to do the majority of I/O from PE0 we are focused on the single client performance to the central pool.



NLCF Center Wide Filesystem

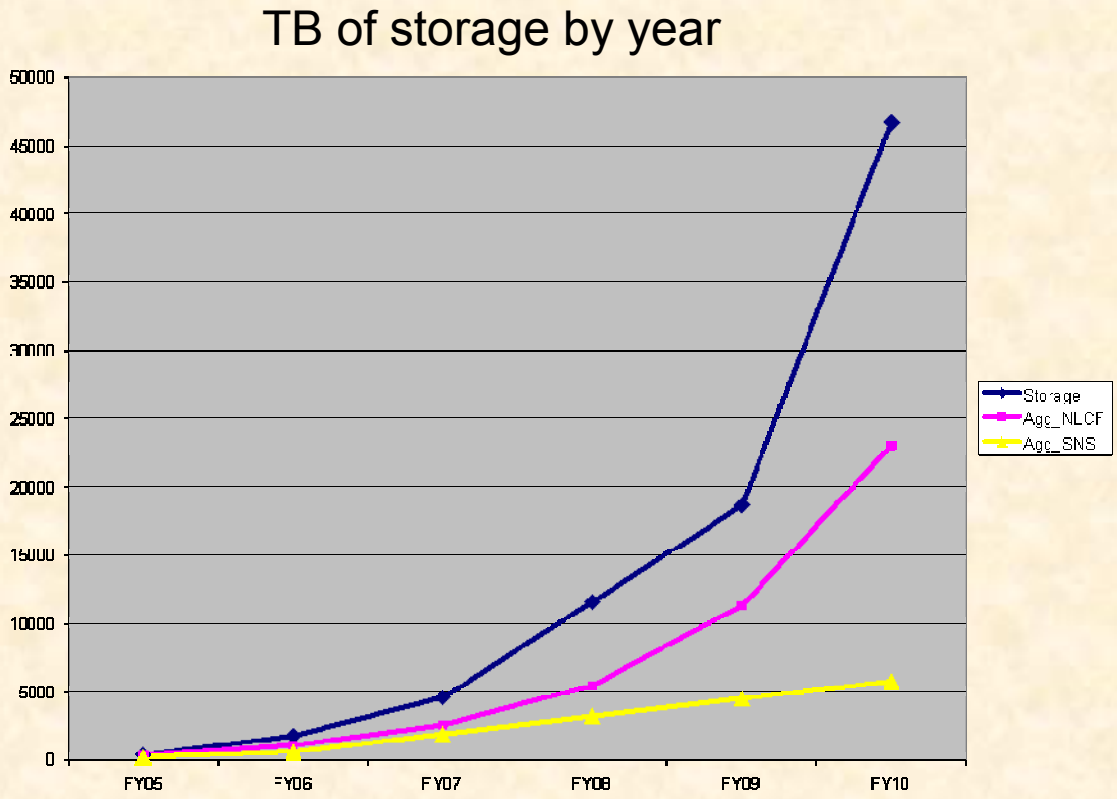
Using Xen to test scalability of Lustre to O(100,000) processors.

SSI Software	**OpenSSI**	**1.9.1**
Filesystem	**Lustre**	**1.4.2**
Basic OS	**Linux**	**2.6.10**
Virtualization	**Xen**	**2.0.2**



Data Management Infrastructure for the Spallation Neutron Source and NLCF

- 50 PB by 2010
- Need to archive, annotate, share, move, replicate
- Current efforts revolve around Lustre, SRB and HPSS
- Connection between database-assisted data management and file-based raw data I/O



FreeLoader: Collaborative Caching for Large Scientific Data

http://www.csm.ornl.gov/~vazhkuda/Morsels

Problem Space:

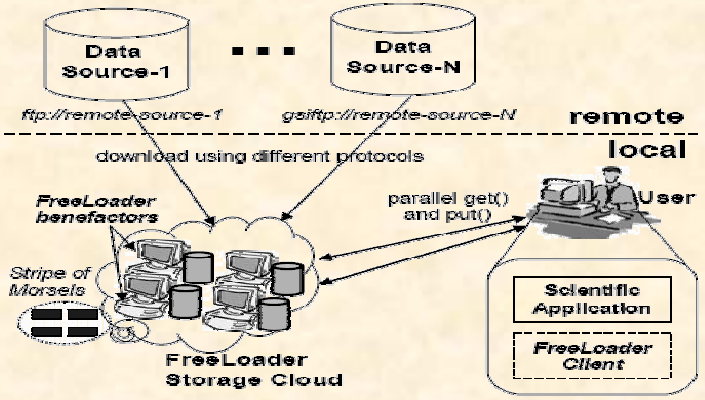
- **Data Deluge:** Increasing dataset sizes (*NIH, SDSS, SNS, TSI)*
- **Locality of interest:** Collaborating scientists routinely analyze and visualize these datasets
- **Desktop, an integral part:** End-user consumes data locally due to ever increasing processing power, convenience & control; But *limited by secondary storage capabilities*

Enabling Trends:

- **Unused Storage:** More than **50% desktop storage unused**
- **Immutable Data:** Data is usually write once read many, with remote source copies
- **Connectivity:** Well connected, secure LAN settings

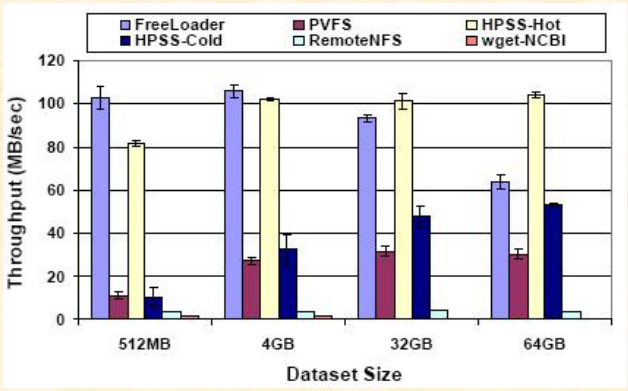
FreeLoader Aggregate Storage Cache:

- Scavenges O(GB) of contributions from desktops
- Parallel I/O environment across loosely-connected workstations, **aggregating I/O as well as network BW**
- NOT a file system, but a low-cost, local storage solution enabling client-side caching and locality



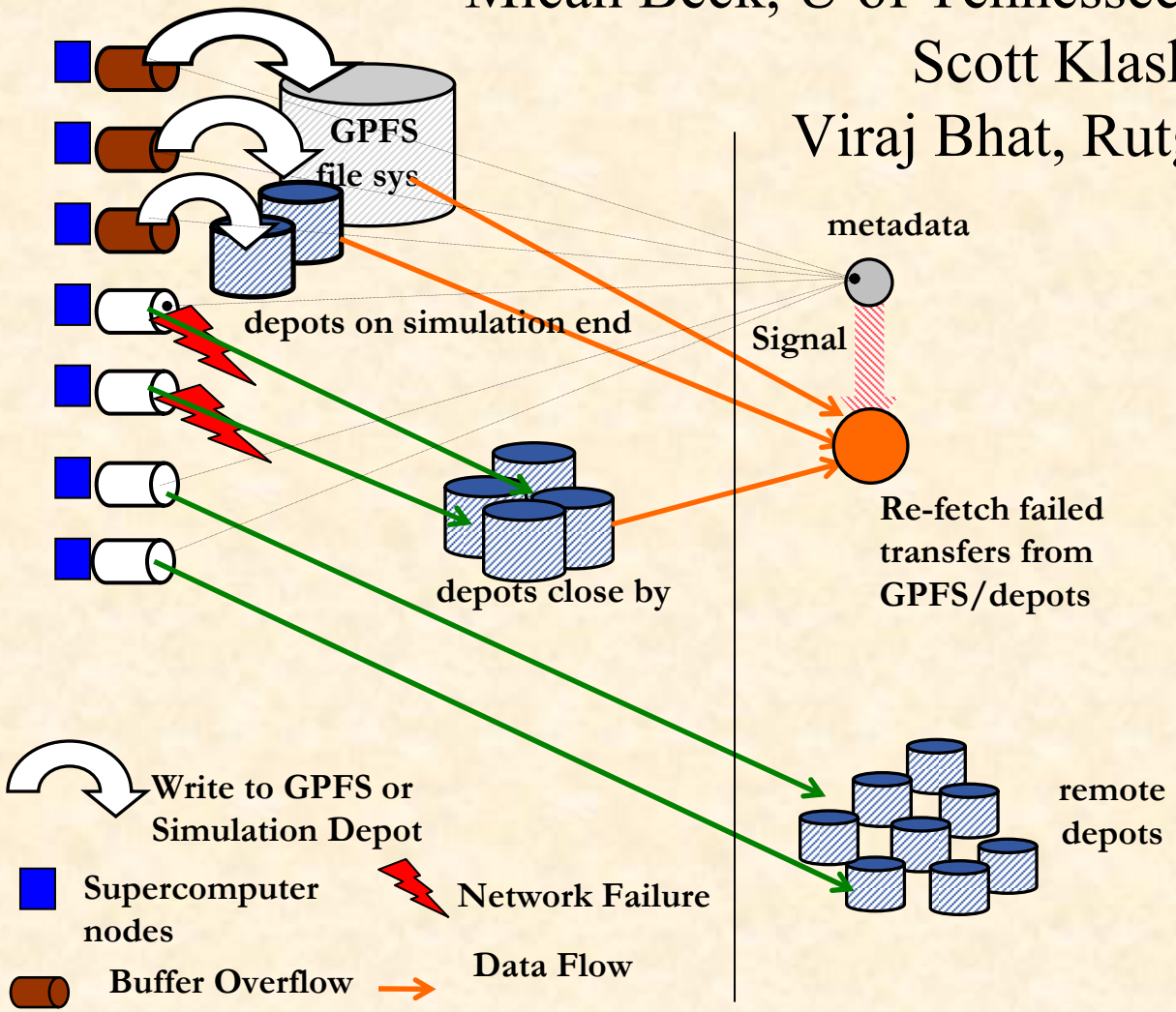
Initial Results:

- Striping across desktops delivers comparable aggregate BW to file systems
- In SC'05



Logistical Networking Used for Failsafe Wide Area Data Streaming

Micah Beck, U of Tennessee & ORNL
Scott Klasky, ORNL
Viraj Bhat, Rutgers Univ.



Future Directions for ORNL storage

- Important to improve serviceability of NLCF high-end storage
  - Center wide file system
    - heterogeneous set of clients
    - Focusing on single client IO performance
    - Aggregate IO performance must scale with # of disks, clients
  - Failover to
    - storage caches or archival storage seamlessly
- Availability
  - Potentially large # of disks in future storage systems
  - Replication based on access patterns of datasets
- Profile NLCF applications' I/O usage
- Track and benchmark I/O subsystems on NLCF platforms

Acknowledgments

- DOE Office of Science
- ORNL LDRD—TCSS Initiative
- DOE Basic Energy Sciences
- NSF-TeraGrid